WEKA

SOLUTION BRIEF

# WEKA Delivers Unmatched Speed for AI and HPC Environments

### Challenges

- Data processing delays caused by high latency limit the speed of insights for data-heavy workloads.
- Insufficient throughput creates bottlenecks in GPU utilization, slowing down AI model training and HPC computations.
- Traditional architectures with complex data paths introduce inefficiencies, delaying time to insight.
- Performance suffers under peak loads, forcing data teams to balance speed against system stability.

### Solution

- WEKA's platform is designed to eliminate latency, maximize data throughput, and fully optimize performance in demanding environments like AI, machine learning, and scientific research, giving data teams a competitive edge in speed and efficiency.

### Benefits

- Gain real-time data access with ultra-low latency to unlock rapid insights and accelerate workflows.
- Maximize GPU utilization up to 93%, enabling faster AI model training and efficient data analysis.
- Reduce time-to-insight for data-driven decision-making across high-throughput applications.
- Maintain peak performance during heavy workloads, ensuring consistent speed and responsiveness.

## Accelerate Data-Intensive Workloads and Optimize Performance

As workloads grow increasingly data-intensive, cloud service providers face mounting challenges in meeting the speed and performance demands of AI and high-performance computing (HPC) environments. Traditional storage systems often fall short, creating delays and bottlenecks that reduce productivity and slow down insights.

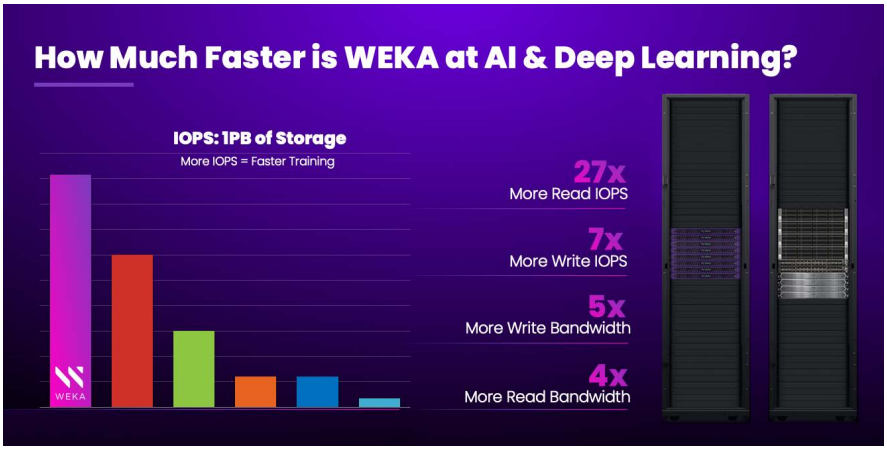
WEKA's high-speed data platform breaks through these limitations, delivering ultra-low latency, rapid throughput, and consistently fast data access. Powered by technologies like kernel bypass to minimize data path latency and NVMe alignment for efficient data transfer, WEKA provides the speed and efficiency essential for handling resource-intensive AI models and large-scale HPC applications. This ensures consistent performance even as demands grow, making WEKA the ideal choice for high-performance environments.

**Ultra-low latency performance:** Lightning-fast data access through kernel bypass and NVMe alignment.

**Maximized GPU Utilization:** Consistent high-speed throughput boosts compute efficiency.

**Faster Time-to-Insight:** Rapid data processing keeps teams ahead.

**High Performance Under Load:** Consistent speed, even in peak workflows.



## Technical Deep Dive

WEKA employs technologies that remove the traditional barriers to throughput and performance. Innovations like **kernel bypass, NVMe storage alignment, and distributed metadata management** ensure that even the most demanding workloads can run with minimal latency and maximum efficiency.

- **Kernel Bypass for Reduced Latency**
  Traditional storage systems rely on the kernel for data processing, which introduces latency and overhead as data moves through multiple layers. WEKA's **kernel bypass** eliminates these bottlenecks by enabling direct data access between the application and storage, reducing I/O overhead and delivering lightning-fast data processing, even under heavy workloads.

- **NVMe Storage Alignment for High-Speed IO Operations**
  WEKA is optimized for **NVMe (Non-Volatile Memory Express) storage**, aligning data at **4K granularity** to maximize I/O efficiency. This tight alignment ensures that every bit of storage is utilized effectively, resulting in faster read/write operations and improved throughput. By bypassing the limitations of traditional storage systems, WEKA enables cloud service providers to achieve **peak performance** for their most data-intensive applications.

- **Zero-Copy Architecture for Accelerated Data Processing**
  WEKA's **zero-copy architecture** minimizes data movement by allowing applications to access data directly without unnecessary copying between buffers. This streamlined data path reduces processing time and latency, enabling faster data retrieval and processing. By simplifying data access, WEKA significantly accelerates performance for high-demand workloads, ensuring rapid time-to-insight.

- **Enhanced GPU Utilization with High-Speed Data Throughput**
  In AI and HPC environments, maximizing **GPU utilization** is critical for performance. WEKA ensures consistent high-speed data throughput to GPUs, preventing data bottlenecks that can lead to idle GPU resources and leading to higher GPU efficiency for faster AI model training and data analysis.

**20x Faster GPU Utilization:** WEKA's optimized data pipeline increases GPU utilization from 30% to over 90%, reducing idle time and maximizing efficiency in AI workloads (from: WEKA Makes GPUs 20x Faster).

**20x Reduction in AI Model Training Time:** WEKA has been demonstrated to reduce model training time from 80 hours to just 4 hours, a 20x improvement, accelerating time-to-insight (from: WEKA Data Platform HPC Solution: Redefining Workloads).

**High IOPS and Bandwidth Efficiency:** WEKA's high-speed networking architecture increases IOPS and supports high per-port bandwidth, optimizing network performance for AI and HPC environments (from: High-Speed Networking with WEKA).

**Ultra-Low Latency with Kernel Bypass:** WEKA's kernel bypass technology reduces latency by enabling direct data access, minimizing bottlenecks, and supporting real-time applications (from: High-Speed Networking with WEKA).

**64% Improvement in Storage Efficiency in the Cloud:** WEKA's platform delivers up to a 64% improvement in storage efficiency, enabling faster data access and processing for high-demand applications (from: The Economic Value of WEKA in the Cloud).

**Up to 2.7x Lower Cost per Job:** WEKA's platform optimizes data processing, reducing the cost per job by up to 2.7x while maintaining high performance (from: The Economic Value of WEKA in the Cloud).

## Use Cases

- **AI and Machine Learning**
  Speed is critical to reduce training times and accelerate inference. WEKA enables faster data processing, allowing AI models to train more quickly and efficiently. By reducing latency and increasing throughput, WEKA helps AI practitioners achieve faster time to insight, giving businesses a competitive edge in developing smarter models faster.
- **High-Performance Computing (HPC)**
  WEKA eliminates HPC bottlenecks with its low-latency, high-throughput architecture, ensuring that HPC workloads are processed efficiently, even as datasets grow. Whether it's scientific research, financial modeling, or engineering simulations, WEKA delivers the performance needed to handle massive datasets without compromising responsiveness.

## Industry Applications

- **Financial Services**
  WEKA delivers ultra-low latency and high throughput to support latency-sensitive applications in financial modeling, high-frequency trading, and risk analysis, enabling faster decision-making and transaction speeds.
- **Life Sciences and Drug Discovery**
  WEKA accelerates AI-driven drug discovery by enabling faster model training and data processing, reducing training times from days to hours. This rapid processing supports critical, time-sensitive applications in oncology, infectious diseases, and biotechnology.
- **Genomics and Healthcare**
  WEKA's high-speed data infrastructure supports large-scale genomic sequencing and analysis workloads, reducing data processing times and enabling rapid insights in areas such as precision medicine and population health studies.
- **Autonomous Vehicle Development**
  WEKA maximizes GPU utilization and speeds up data processing for autonomous driving workloads, supporting faster model training and enabling real-time decision-making essential for autonomous systems.
- **Media and Entertainment**
  WEKA supports high-performance media workloads, delivering rapid data processing and low latency for video rendering, editing, and effects, allowing creative teams to complete projects faster and more efficiently.
- **Academic and Research Institutions**
  WEKA enhances research capabilities by providing high-throughput data access and low latency for complex simulations and modeling. This performance accelerates research outcomes in disciplines ranging from economics to scientific computing.
- **Enterprise AI and Language Models**
  WEKA accelerates large-scale AI model training and deployment, reducing time to production for enterprise language models by minimizing latency and maximizing throughput, which enables faster, tailored AI solutions for enterprise applications.
- **Generative AI and Model Training**
  WEKA accelerates generative AI model training by reducing latency and optimizing GPU utilization, enabling faster time-to-insight for resource-intensive models across imaging, video, 3D, and natural language applications.

## Why WEKA?

- **Accelerated Time to Insights:** WEKA's high-speed data processing allows organizations to transform raw data into actionable insights faster, enabling quicker decision-making and maintaining a competitive edge.
- **Rapid Model Training:** WEKA's low-latency, high-throughput platform significantly reduces model training times, accelerating the development cycle for AI and machine learning applications and shortening time to market.
- **Faster Time to First Token for Inferencing:** Inference workloads for autonomous vehicles, fraud detection, recommendation engines, and others demand instant data access. WEKA delivers millions of IOPS with ultra-low latency, enabling models to produce outcomes faster.
- **Optimized for Ultra-Fast Data Access:** WEKA's advanced architecture—including kernel bypass and NVMe alignment—ensures minimal latency and maximum throughput, delivering lightning-fast data access and real-time responsiveness, even during peak workloads.
- **Enhanced GPU Utilization:** By sustaining high data throughput, WEKA maximizes GPU efficiency, ensuring computational resources are fully leveraged for intensive AI and HPC workloads.
- **Consistent High-Speed Performance at Scale:** WEKA's distributed architecture scales seamlessly with workloads, providing high-speed performance even as data volumes and user demands grow, ensuring no compromise in speed or efficiency.
- **High-Throughput Data Streaming for Real-Time Applications:** WEKA's architecture supports continuous data streaming, ideal for real-time applications such as AI inference, financial transactions, and autonomous decision-making where immediate data processing is essential.





WKA414-01 12/24